基于本体的国史知识检索平台构建研究[1]

王颖[1]　张智雄[1]　孙辉[2]　雷枫[2]
[1](中国科学院文献情报中心　北京　100190)
[2](当代中国研究所　北京　100009)

摘要：[目的/意义] 构建国史知识检索平台，提高用户获取国史知识的效率，促进国史宣传和教育。[方法/过程] 提出基于本体的国史知识检索平台构建思路与总体框架，在构建国史本体知识库的基础上，采用 Neo4j 数据库作为 RDF 数据仓储，创建基于 Solr 的实例索引、三元组索引和词条索引，针对 多种检索需求设计实现了检索引擎的执行流程、检索式构造方法以及查询处理算法，并为国史知识展示设计了可视化实现方式。[结果/结论] 构建了国史知识检索平台，提供实体检索、查询问答、关联检索、时序检索及语义资源浏览等检索与浏览服务，该平台框架及关键技术实现方案为面向领域知识的深度检索服务提供了重要参考。

关键词：本体　实体检索　查询问答　关联检索　可视化

分类号：G250.7

1. 引言

随着互联网的广泛普及，国史宣传教育网站成为了开展国史宣传和国史教育的重要渠道。传统门户网站如“中华人民共和国史教育网”通过栏目导航、网页浏览、多媒体等方式展现新中国成立以来的奋斗历程、伟大成就和成功经验。尽管网站中国史信息资源丰富，但对于国史学习者而言信息量庞大，即使通过全文检索的方式过滤信息，仍然无法直接获得需要的国史知识。为此，有必要构建一个国史知识的检索与浏览平台，进一步提高“中华人民共和国史教育网”的知识性、可读性和互动性，让读者和用户直观地学习和了解国史知识，达到以史鉴今、资政育人的目的。

近几年，国外机构推出了 Kngine、WolframAlpha 等新型的知识引擎系统。Kngine 能够为用户提供更有意义的知识搜索结果，如理解关键词或概念的语义信息，回答用户的问题，发现关键词或概念之间的关系以及链接不同的数据等[1]。WolframAlpha 针对问题可以直接给出有效答案，如在被问到“珠穆朗玛峰有多高”之类的问题时，WolframAlpha 不仅能给出海拔高度，还能显示这座世界第一高峰的地理位置、附近有什么城镇，以及一系列图表[2]。为了让用户能够更快更简单地获得查询信息，传统搜索引擎公司 Google、百度、搜狗等也在逐渐从搜索信息向搜索知识转型，Google 推出了 Knowledge Graph 功能，可以更好的理解用户搜索的信息，并直接将检索词的相关信息呈现在搜索页面中，免去了用户访问信息出处网站这一过程[3]。百度使用框计算技术开发了一些实体搜索和关联推荐的功能，例如搜索“类似盗梦空间的电影”。搜狗的知立方搜索能够进行查询语义理解，通过推理获得“姚明太太的身高”为“190cm”，同时也相应的给出了姚明太太叶莉的资料介绍以及姚明的关系图谱。实现这些创新应用的基础是构建包含实体和相关事实的大规模知识库。如 Google 从 Freebase、维基百科或

---

[1]本文系中国社会科学院哲学社会科学创新工程信息化项目“中华人民共和国史教育网”的研究成果之一

作者简介：王颖(ORCID：0000-0002-1941-3134)，馆员，博士，E-mail：wangying@mail.las.ac.cn；张智雄(ORCID：0000-0003-1596-7487)，研究馆员，博士生导师；孙辉，副编审；雷枫，高级工程师。

全球概览中获得专业的信息构建 Knowledge Graph，2012 年 Knowledge Graph 包含的实体数量就已经超过 5.7 亿个[3]。搜狗知立方对半结构化网页数据进行信息抽取，从文本数据中抽取实体和属性，再联合结构化数据进行异构数据整合，构建了知立方本体知识库[4]。

不仅如此，目前互联网正从仅包含网页和网页之间超链接的文档万维网(Document Web)转变成包含大量描述各种实体和实体之间丰富关系的数据万维网(Data Web)[5]。并且大量的语义化数据如 RDF 和 OWL 仓储数据被发布，开放关联数据（LOD）云的规模日益庞大，在此基础上能够构建更智能的检索应用。如 Sindice 搜索引擎利用爬虫将语义网上的 RDF 数据收集起来，提供了关联数据网的实体检索和查询服务[6]。Semplore 搜索引擎使用关键词检索和结构化检索的混合检索机制，创建了关键词、概念和关系三种类型的倒排索引，提供关联数据的检索查询、分面导航服务[7]。FREyA 是一个面向本体的交互式自然语言查询接口，使用语法分析和基于本体的查询对用户问题进行解释，利用用户反馈消解歧义，再构造 SPARQL 查询完成查询应答[8]。Treo 集成实体搜索、扩散激活搜索和基于 Wikipedia 的语义相关度计算对关联数据网进行检索，将解析的用户查询与数据集中的数据表示进行语义匹配[9]。MEANS 是一个结合自然语言处理和语义网技术的医学问答系统，它使用自然语言处理技术对医学问题和文档进行深层分析，为文档建立 RDF 标注，将用户问题转换为 SPARQL 查询，实现针对医学文档集的查询问答[10]。

在借鉴现有相关研究和系统的基础上，本文提出了基于本体的国史知识检索平台的建设思路，介绍了平台的总体框架，对关键技术实现方案进行了阐述，最后展示了平台的实现效果。

2.平台设计

2.1 设计思路

国史知识检索平台的建设目标是在构建的国史本体知识库基础上，对外提供检索查询、浏览导航、知识展示等功能，让用户可以方便地阅读收录和编辑的国史知识。为达到这个目标，我们提出了平台的设计思路：

（1）建立国史本体知识库的有效存储和灵活访问机制，实现国史知识的语义组织和进一步利用；

（2）借助本体的结构化语义和推理规则，实现国史知识的细粒度揭示和潜在知识挖掘；

（3）提供面向国史知识实体的检索与浏览功能，支持实体之间的关联发现服务；

（4）实现查询问答功能，允许用户提出国史问题，系统直接返回答案，而不是相关的历史文本资料；

（5）通过国史知识图谱的方式可视化展示国史知识关联，便于用户直观地了解相关信息；

（6）通过文本资源的语义链接进行知识实体的跳转浏览，允许用户进行延伸性阅读。

2.2 总体框架

根据平台设计思路，确定国史知识检索平台的总体框架如图 1 所示，包括数据层、功能层和服务层三个层次。数据层提供数据仓储和数据访问等基本功能，底层分别存储了构建的国史本体知识库以及收集的国史教育文本资源，并为支持

检索应用构建了实例索引、三元组索引和词条索引，同时提供对底层存储以及索引的数据调用接口，允许检索引擎进行数据访问。功能层由国史知识检索引擎完成检索平台的查询分析、检索调度、结果计算等核心功能。检索引擎对用户输入进行查询解析，推荐相关检索词，并根据解析结果构建相应的检索表达式进而执行检索调度完成检索任务。检索过程中利用了国史本体的推理机制挖掘隐含知识来扩展检索的范围，并对返回结果进行统计和排序等操作。在服务层，为用户提供实体检索、查询问答、关联检索、全文检索和时序检索等基本检索服务，由检索引擎执行检索，通过可视化显示国史实体的知识图谱以及文本资源语义化浏览等提高用户对于国史知识的获取效率，并借助平台揭示潜在知识和推导隐含知识，支持用户进一步的知识发现。

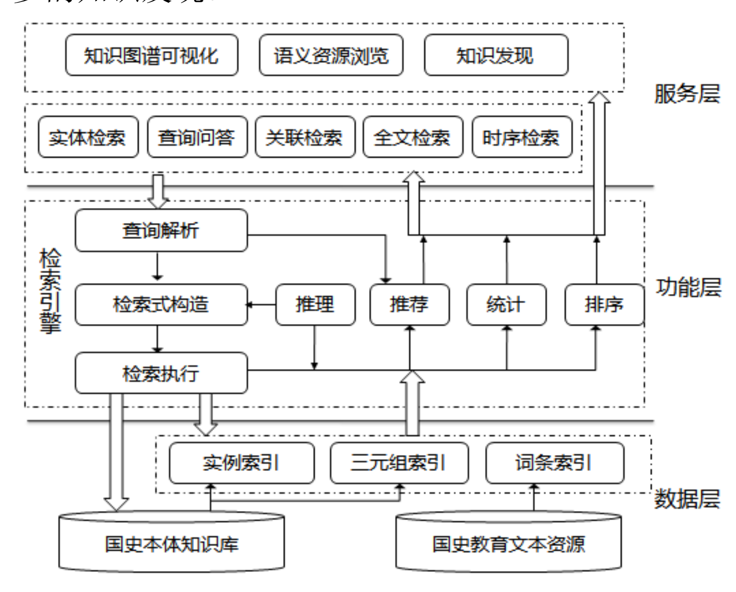


图 1 总体框架

3.关键技术实现

3.1 国史本体知识库构建

国史本体知识库是建立国史知识检索平台的基础。在借鉴其他历史本体构建经验的基础上，针对国史领域特色，作者提出了国史本体的思路和方法，构建了国史本体的概念模型[11]。通过国史本体对人物、机构、会议、事件等知识实体、概念及其关系进行规范化和语义化表示。国史本体定义了事件、会议、人物、机构、文件、理念或术语等 19 个类以及 20 个数值属性、76 个对象属性，并根据常识知识定义了一些属性约束和推理规则。从而确定了国史知识实体的分类以及实体关系类型，用于指导国史本体知识库的构建。

国史本体知识库的主要知识来源是辅助国史教育的基础工具书，如《中华人民共和国国史百科全书》、《中国共产党历史大辞典》等。由于单纯依靠人工创建国史知识实体和关系耗时耗力，为此使用自动处理与人工加工相结合的方法从文本中提取明确的国史知识。首先，由领域专家挑选重要的国史词条作为国史知识发掘的基础资源，收集整理国史相关的人物表、机构表等主题词表，并从一些工具书中提取词条名称、类型、时期等重要的元数据，通过人工整理和校验，转化为基础的本体实例以及基础数值属性。利用这些实例数据对收集的国史词条进行自动标注，提取国史实体可能相关的关系或属性数据，作为人工加工的辅助信息。在国史本体概念模型的约束下，人工修订或构建本体实例，编辑实例属性和关系

信息，并对编辑的数据进行审核和管理，逐步构建出包含重要国史知识的国史本体知识库。目前构建的国史本体知识库共包括 15,602 个实例，5,147 条属性信息，21,503 条实例关系。

3.2 neo4j 存储

针对国史本体知识库的特点，选用图数据库 Neo4j 作为底层数据仓储，支持结构化检索、关联检索等复杂的检索需求。Neo4j 是一个高性能的 NoSQL 图数据库，用高效的图数据结构代替传统的表设计，将数据保存为图的节点以及节点之间的关系，并提供了图的查找和遍历功能。图数据库 Neo4j 具有很好的可扩展性和灵活性，适用于复杂关系的管理与查询推理，符合基于 RDF 图数据模型的本体知识库三元组存储和 SPARQL 查询需求。

在 Neo4j 数据库中，将国史本体知识库中的每个实例保存为一个节点，实例的数值属性存储为节点的属性和属性值，实例关系存储为节点和节点之间的二元有向关系。例如，存储实例“毛泽东”时，在 Neo4j 数据库中新增一个节点，自动生成节点 ID 为 327，根据该实例在知识库中的信息，定义节点属性“label”的数学值为“毛泽东”，定义节点属性“altLabel”的属性值为“毛主席”。此外，依据国史本体定义“曾任职务”、“所属党派”等关系类型，定义节点关系 relationship[328]：曾任职务（“毛泽东”，“中华人民共和国主席 1949”），relationship[326]：所属党派（“毛泽东”，“中国共产党”）等。

3.3 索引设计

采用 Apache Solr 索引技术，对国史本体知识库中的实例和属性关系以及国史教育文本资源构建索引，提高查询速度和系统响应时间。根据检索需求，构建了三种索引包括：

（1）实例索引

实例索引用于实现国史实体的快速检索和分面导航等功能，主要字段包括：id、规范名称 label、其他名称 altLabel、首字母 initial、实体类型 entityType、相关词条 textItemID 等。检索引擎在用户输入检索词时根据用户的输入实时检索实例索引给出检索词提示，对用户输入解析后利用 label 和 altLabel 字段进行精确匹配和模糊匹配，返回匹配或推荐的实例，并根据实体类型和首字母等索引字段对实例进行分面导航。

（2）三元组索引

三元组索引用于支持对国史本体知识库的快速检索，即对知识库的三元组构建索引，主要字段包括三元组陈述主体的 sID、名称 sLabel、类型 sType、类型名称 sTypeValue，谓词的 pID、名称 pLabel、类型 pType、类型名称 pTypeValue 以及客体的 sID、名称 sLabel、类型 sType、类型名称 sTypeValue、知识来源词条 textItemID 等。三元组索引的构建为结构化的关系检索提供了前提条件，检索引擎构建类似 SPARQL 查询语言的 Solr 检索式实现国史本体知识库的检索应用。

（3）词条索引

词条索引针对收录的国史教育资源文本词条进行索引，支持文本资源的全文检索功能，主要字段包括：词条 ID、题目 itemTitle、文本内容 itemText、类型 itemType 以及词条来源库 itemSource 等。

3.4 检索引擎功能实现

为满足用户不同的检索需求，国史知识检索平台设计了三个检索入口：普通检索、关联检索和时序检索。普通检索允许用户输入检索词或语句查询相关的国史知识，关联检索允许用户输入两个检索词来获得实体之间的相互关联，

时序检索用于检索限定时间范围内的事件、会议、文献等。

3.4.1 检索执行流程

面向三个检索入口，检索引擎分别依据检索输入类型、检索目标范围、输出结果等设定相应的检索处理方式，具体执行流程如图 2 所示。

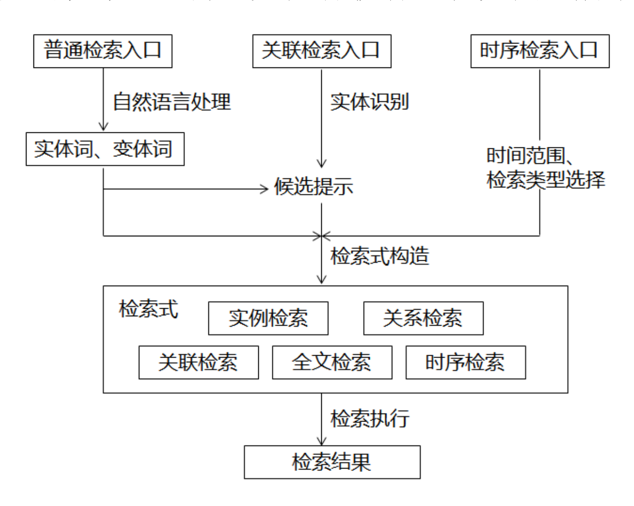


图 2 检索执行流程

普通检索入口允许用户输入任意的检索词查询目标实体的相关国史知识，同时也可以输入查询问题获得国史知识答案。为此，检索引擎首先使用自然语义处理技术对普通检索输入进行处理，从检索词或语句中提取国史本体知识库中的实体词和预定义的变体词，依据命中情况构造实例检索式、全文检索式、关系检索式或关联检索式。如果存在具有相同名称的实体或无命中结果将候选实体和模糊匹配推荐的实体提示给用户，由用户选择检索目标，再构造相应的检索式。

关联检索要求用户输入两个实体的名称。同样，如果通过实体识别命中知识库中两个不同的实体则直接构建关联检索式进而执行检索，如果命中一个相同的实体，则提示给用户转化为实体检索，如果命中多个或无命中结果则将候选实体和模糊匹配推荐的实体返回给用户，由用户选择两个实体再构建检索式。

时序检索由用户输入以年份为单位的起始时间和终止时间范围，选择事件、会议或文献等返回类型，检索引擎根据用户的选择构造出相应的检索式。

检索式构造完成后，由检索引擎选择具体的索引或数据库执行检索任务，返回检索结果，并对检索结果进行统计和排序。

3.4.2 检索式构造

国史本体知识库的构建使得国史知识成为了可计算的结构化数据，使用类似 SPARQL 语言的查询方式对知识库进行检索可实现细粒度国史知识的检索服务。经过查询解析，检索引擎根据检索需求构建如下几种不同的检索表达式，分别执行检索任务。

（1）实例检索

实例检索通过实例名称、其他名称、类型等字段检索实例索引，返回精确匹配或模糊匹配的实例结果，并对结果集进行名称排序或相关度等操作。例如检索精确匹配名称为“一国两制”的实例检索式为：

query=(label: "一国两制") OR (altLabel: "一国两制") (1)

检索引擎将检索式传递给实例索引服务器执行检索任务，然后对检索结果进行封装和处理输送到前端显示。

（2）关系检索

关系检索主要面向三元组索引。如果已知国史本体知识库中某个实例查询其相关属性和关系，即通过指定三元组（s,p,o）中 s 或 o 为指定实例，返回三元组结果集。例如用户在普通检索入口输入检索词“毛泽东”，检索引擎首先进行查询解析，获取命中实体对应的实例 ID 为 6787，检索引擎构造关系检索式为：

query=(sID:6787) OR (oID:6787) (2)

如果已知国史本体知识库中某个实例和指定属性查询该实例通过指定关系关联的实例或具有的属性值，即通过指定三元组（s,p,o）中 s 和 p 或者 p 和 o 来查询知识库返回对应的三元组结果集。例如查询“中共十一届三中全会”的“参会者”，检索引擎针对三元组索引构造关系检索式为：

query= (sLabel:"中共十一届三中全会") AND (pLabel:"参会者") (3)

根据领域特性，国史本体中定义了一些属性约束和推理规则，在检索式构造时，检索引擎应用本体推理机制对检索式进行了重构。如上例中，由于本体中定义了对象属性“参会者”和“参会”相互为逆关系，以及“会议的发言人和报告人一定参加了会议”的规则，检索式重构为：

query=((sLabel:"中共十一届三中全会") AND (pLabel:"参会者"))
OR ((pLabel:"参会") AND (oLabel:"中共十一届三中全会"))
OR ((sLabel:"中共十一届三中全会") AND (pLabel:"发言人或报告人"))
(4)

从而发现未明确表示的国史知识，扩展了关系检索的范围。

（3）关联检索

关联检索针对国史本体知识库中三元组集合所形成的图结构特性，在已知两个不同实例的情况下，通过查询两个实例对应图节点之间的路径获取实例之间直接或间接关系。系统设定关联度选项“近”、“稍远”和“远”，由用户选择其一查询两个实例之间长度不大于 2、3 或 4 的路径，返回经过路径的三元组集合。关联检索的执行借助 Neo4j 数据库的图遍历机制，通过 Cypher 查询语句检索 Neo4j 数据库获取查询结果。例如检索“邓小平”和“中共十一届三中全会”之间“稍远”的关联，检索引擎首先对检索词进行实体识别，获取命中实体 ID 为 5904 和 14563，则构造 Cypher 查询语句：

```
start a=node(*), b=node(*) match p=a-[*0..3]-b
     where a.source_id=5904 and b.source_id=14563
     return p order by length(p) asc;
```
(5)

检索引擎连接 Neo4j 数据库读取查询结果，解析结果集输出到系统前台。

（4）全文检索

全文检索主要针对国史教育文本资源，利用检索词在词条索引中对词条题目和词条内容执行全文检索，通过 Solr 集成 mmseg4j 中文分词工具完成全文检索任务，并根据相关度排序返回相关词条。如检索“中共十一届三中全会”的相关词条，则构造检索式：

query= (itemTitle:中共十一届三中全会) OR (itemText:中共十一届三中全会) (6)

（5）时序检索

时序检索主要与时间类相关，如检索时间区间内发生的事件、召开的会议、出版的文献等。即查询（s,p,o）中 s 的类型为事件、会议或文献（文件、著作、报纸刊物、报告讲话等）而 o 的类型为时间类的三元组。由于通常史料记载的时间取值相对模糊，在构建本体实例时直接保留了原有取值，为了便于计算检索平

台对时间类定义了开始时间和结束时间属性，将时间单位设定到“月”，在构建索引时制定转换规则将“上半年”、“春”、“年底”等模糊时间映射到具体的年月。如查询 1949 年-1950 年间发生的事件，检索式为：

query=(sTypeValue:"事件") AND (oTypeValue:"时间类")
AND (startDate:[194901 TO 195012]) (7)

3.4.3 查询处理算法

普通检索入口允许用户输入检索词和提问语句，其查询处理方法相对关联检索和时序检索而言较为复杂，对此本文提出了查询处理算法。检索引擎首先对用户输入的字符串进行解析。由于国史知识实体的名称通常为中文并且字符长度都不小于 2，因此判断输入字符数目如果小于 2，则执行实例检索，推荐一些候选实体给用户。否则，采用自然语言处理技术，利用知识库中的实例名称和别称作为词典对输入进行实体标注，如果命中的实体数目多于 2 个，则转换为针对文本词条的全文检索，查找与输入相关的词条。如果实体数目为 2 个，则转换为查询两个实体之间关联的检索。否则，利用人工定义的变体词表对输入进行识别，如果没有命中变体词，而命中了一个实体则将查询转换为针对其实体的关系检索返回实体所有相关三元组，或者如果实体和变体词都没有命中则进行实例检索推荐相似的实体给用户。如果命中变体词，则根据预定义关系表将变体词映射到具体的本体属性或类上。例如，将“什么时候召开”、“何时召开”和“召开的时间”映射到本体属性“召开时间”，“地区”、“地方”和“哪里”对应本体类“国家和地区”。再根据属性列表和类列表进一步分析，在命中实体数为 1 的情况下，如果属性列表数目大于零，则将检索转换为针对该实体指定属性的关系检索，否则根据类列表返回命中实体与指定所属类实体之间三元组的关系检索，如果属性列表和类列表都为空则转换为实体的三元组关系检索。在没有命中实体却命中变体词的情况下，如果变体词对应类列表规模大于 0，则转换为指定实体所属类的实例检索，否则转换为输入字符串的全文检索。由此，利用实体词表和变体词表实现了对于查询问题的自然语言处理，形成针对具体情况的多种检索式，达到了从知识库和文本资源中查找国史知识的目的。具体算法如下所示：

表 1 查询处理算法

```
01.输入：用户输入字符串 input
02. IF 输入中文字符长度 length(input)<2
03.     推荐相关实例给用户 query=(实例检索,(label:input OR altLabel:input));
04. ELSE{
05.     获取命中实体列表 entityList=entityAnnotation(input);
06.     IF 实体数目 size(entityList)>2
07.         转换为全文检索 query=(全文检索,input);
08.     ELSE{
09.         IF 实体数目 size(entityList)=2
10.             转换为关联检索 query=(关联检索,entity1,entity2);
11.         ELSE{
12.             去除已标注字符 input_new=replace(input);
13.             IF 中文字符长度 length(input_new)<2
14.                 变体词列表为空 variantList=null;
15.             ELSE{
16.                 获取命中变体词列表 variantList=variantAnnotation(input);
17.                 IF 变体词数目 size(variantList)=0 {
18.                     IF 实体数目 size(entityList)=1
19.                         转换为关系检索 query=(关系检索,(s:entity1));
20.                     ELSE
21.                         推荐相关实例给用户 query=(实例检索,(label:input OR altLabel:input));
```

```
            } ELSE{
                通过变体词表获得变体词对应的属性列表 propertyList 和类列表 classList;
                IF 实体数目 size(entityList)=1{
                    IF 对应属性数目 size(propertyList)>0
                        转换为关系检索 query=(关系检索,(s:entity1)AND(p in propertyList));
                    ELSE{
                        IF 对应类数目 size(classList)>0
                            转换为关系检索 query=(关系检索,(s:entity1)AND(o_type in classList));
                        ELSE
                            转换为关系检索 query=(关系检索,(s:entity1));
                    }
                } ELSE{
                    IF 对应类数目 size(classList)>0
                        转换为实例检索 query=(实例检索,(entity in classList));
                    ELSE
                        转换为全文检索 query=(全文检索,input);
                }
            }
        }
    }
}
}
输出：query
```

3.5 可视化实现

为直观地向用户展示国史知识之间的相关关联，选用 Cytoscape Web 工具实现国史知识图谱的可视化展示。Cytoscape Web 是一个开源的图形可视化库，它的数据模型支持节点和有向边，可以定义节点和边的名称和类型，满足知识库实例和关系名称显示和配置的应用需求。Cytoscape Web 网络显示的主要组件通过 Flex/ActionScript 实现，可视化样式多样，并且提供了实现网络视图的定制和交互的 JavaScript API，支持视图的缩放、拖拽、节点与边的点击事件、类型筛选等功能。

在项目工程中引入 Cytoscape Web 的 js 文件，通过后台程序封装可视化显示的数据信息，包括节点和边的颜色、名称、类型、显示样式等信息，生成 json 格式传递至前台显示，利用自定义 jQuery 方法进行解析，配置图形样式并实现 Ajax 操作，通过引用 Cytoscape Web 的 js 方法绘制图形和实现事件响应。

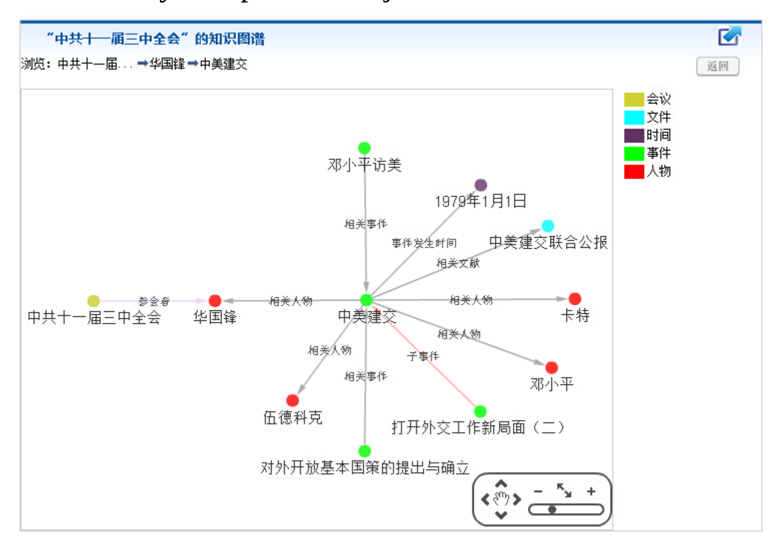


图 3 可视化展示

在可视化窗口中，用节点代表实体，两个节点之间的有向边揭示实体之间的关系，通过不同的颜色代表实体所归属的类，直观地呈现知识图谱。同时，提供丰富的用户交互功能，例如通过浮动的操作面板上下左右移动、缩小或放大图形以及适应窗口尺寸。在空白处点击几秒钟鼠标箭头变成小手图标可以实现对图形的整体拖动，同样也可以拖拽节点。左键单击节点可以进一步浏览该节点的知识图谱，右键单击节点显示节点的详细信息，左键单击边可以在图上显示该边的名称，右键点击边可以查看该节点关系的知识来源（知识来源文本条目），点击窗口右侧本体类颜色注释可进行筛选窗口内相关实体的类型。系统支持可视化浏览功能，用户可以通过点击知识图谱中的节点不断获取相关的国史知识。如图 3 所示，如浏览“中共十一届三中全会”的知识图谱时，显示“中共十一届三中全会”的参会者，点击其中一个参会者人物“华国锋”，可以进一步浏览它的相关知识，获得事件“中美建交”通过“相关人物”关系指向“华国锋”，点击“中美建交”节点进一步获得相关的人物、事件、文献、子事件等知识。系统保留了浏览的路径和历史记录，允许返回操作。由此实现了基于知识图谱的漫游式浏览，有利于用户直接、便捷地获取国史知识。

4 平台实现效果

在上述平台设计和关键技术实现的基础上，完成了国史知识检索平台的建设。该平台是基于 B/S 模式的 Web 应用系统，使用 Java 语言开发，采用 springMVC 和 hibernate 作为开发框架，数据库使用 Neo4j 2.1.2 Win64 免费版本，Solr 版本为 4.7.2，系统运行环境为 Windows 2008 服务器操作系统，以 Tomcat 6.0.4 作为 Web 服务器，使用 JDK 1.7。检索平台实现了实体检索、查询问答、关联检索、时序检索、语义资源浏览等服务功能。

4.1 实体检索

区别于将检索词匹配的国史教育文本资源呈现给用户的传统检索方式，国史知识检索平台将国史本体知识库内部与检索目标匹配的实体相关知识通过可视化方式展示给用户。用户不需要通过阅读文本信息就可以直观地了解相关的国史知识，同时通过知识图谱的点击操作可以进行延展性阅读，这使得国史知识的获取更有效率。如检索事件“土地改革运动”返回的知识图谱如图 4 所示，可以清晰的看到与“土地改革运动”相关的会议、事件、文件、理念或术语、人物、机构等。同时检索平台返回“土地改革运动”的相关资料，用户可以继续阅读相关国史教育文本资源。

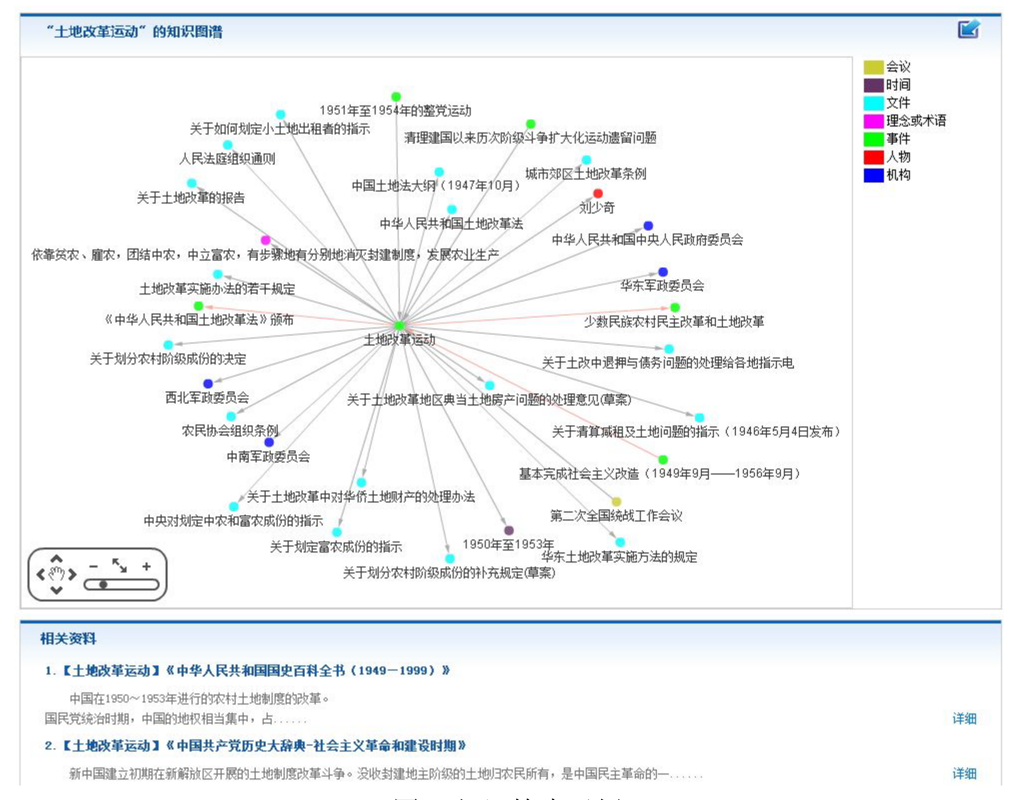


图 4 知识检索示例

4.2 查询问答

为满足用户使用自然语言提问的检索需求，设计和实现了国史知识查询问答功能。利用自然语言处理技术对用户提出的问题进行分析，构造针对国史本体知识库的结构化检索式，返回结果知识图谱。例如，用户输入“谁提出了‘帝国主义和一切反动派都是纸老虎’”，返回如图 5 所示的检索结果知识图谱，自动返回“帝国主义和一切反动派都是纸老虎”的“理念提出者”是“毛泽东”。

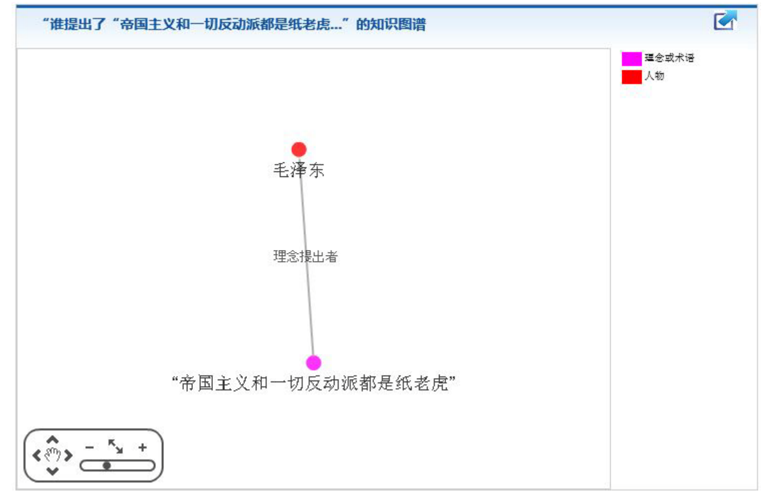


图 5 查询问答示例

4.3 关联检索

关联检索借助国史本体知识库的图结构发现实体之间的相互关联，获取子图结构，挖掘潜在的知识。例如检索“陈云”和“两个凡是”的“近”的关联，可获得如图 6 所示的知识图谱。从图中我们可以了解到：“陈云”作为“发言人或报告人”与“中共中央工作会议（1977 年 3 月）”会议相关联，该会议的“相关概念

或术语”为“两个凡是”，事件“邓小平第三次复出”和“党内外对‘两个凡是’的批评和抵制”都与人物“陈云”、概念或术语“两个凡是”相关，由此通过子图展示了“陈云”和“两个凡是”的内在联系。

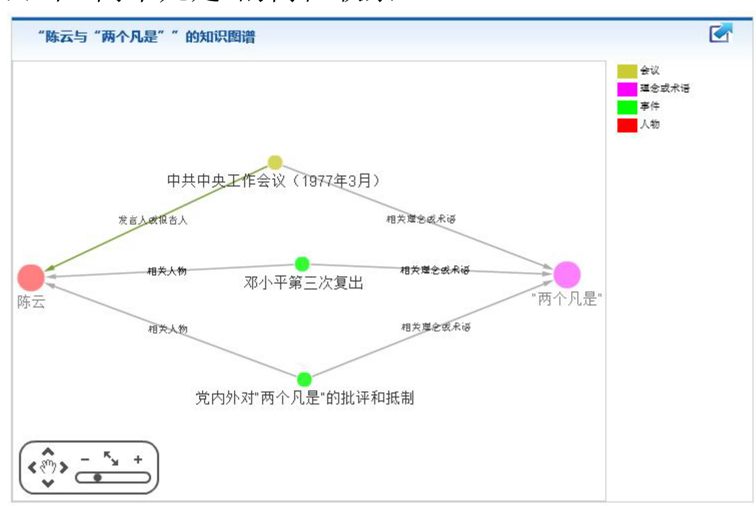


图 6 关联检索示例

4.4 时序检索

时序检索允许用于选择时间范围，查找这段时间的事件、会议、文献等信息。检索 1949 年到 1950 年的事件和会议，获得如图 7 所示的结果，包括 67 个时间和 37 个会议，按照时间顺序排列，点击实体名称可以进一步浏览它的知识图谱。



图 7 时序检索示例

4.5 语义资源浏览

检索平台在词条详细信息页面提供了语义资源浏览功能，通过文本标注程序将文本中重要实体标识出来，利用不同的颜色区分不同的类型。点击标注的实体名称自动跳转到该实体的知识图谱页面上，方便用户通过知识图谱进一步了解相关国史知识。如图 8 所示，“土地改革运动”词条中标注了任务、会议、事件、机构、国家和地区、特殊群体、概念或术语和文件等国史实体或概念。



图 8 语义资源浏览示例

5. 结论

本文以国史本体为基础，构建了国史知识检索平台，探讨了本体知识库的存储、索引、可视化以及基于本体的知识检索技术。本研究利用本体从更精细的角度来表示和组织国史知识，实现了实体检索、查询问答、关联检索和时序检索等结构化、细粒度的检索服务，同时支持了国史知识的深度挖掘与探索，扩展了信息检索的深度，提高了知识获取的效率。此外，通过知识图谱可视化和语义资源浏览丰富了检索结果的呈现形式，改进了用户体验，提高了“中华人民共和国史教育网”的互动性和新颖性。在后续的工作中，将继续增加国史本体知识库的实例，丰富知识库的内容，同时提高检索平台在大规模数据处理上的性能，进一步提升平台服务效果。

参考文献：

[1] Kngine[EB/OL]. [2015-03-10]. http://www.baike.com/wiki/Kngine.

[2] WolframAlpha[EB/OL]. [2015-05-10] http://baike.baidu.com/link?url=7hCseBbIipm9Xo0xOeMo7jVuCrgcUMWTO7F-yW7lUr9YkjUw7_WmG3_i9yeyiRAoRVCvNCPbD6JxzNfyJ4ET6_.

[3] Singhal, Amit . Introducing the Knowledge Graph: Things, Not Strings[EB/OL]. [2013-04-10]. http://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html.

[4] 张阔. 从搜索信息到搜索知识——技术架构[EB/OL].[2013-03-26]. http://weibo.com/1870490225/zbzDwq5TF#_rnd1435219297630.

[5] 王昊奋. 大规模知识图谱技术[EB/OL]. [2014-10-12]. http://blog.sciencenet.cn/home.php/fgcfmt/home.php?mod=space&uid=1225851&do=blog&id=801901.

[6] Tummarello G, Delbru R, Oren E. Sindice.com: Weaving the Open Linked Data[C]. Proceedings of 6th International Semantic Web Conference, Springer. LNCS 4825, 2007: 552-565.

[7] Wang H. et al. Semplore: A Scalable IR Approach to Search the Web of Data [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2009, 7(3): 177-188.

[8] Damljanovic D, Agatonovic M, Cunningham H. FREyA: An Interactive Way of Querying

Linked Data Using Natural Language[C] Proceedings of ESWC 2011 Workshops, LNCS 7117, 2012: 125-138.
[9] Freitas A. et al. Querying Linked Data Using Semantic Relatedness: A Vocabulary Independent Approach[C]. Proc. 16th Int'l Conf. Applications of Natural Language to Information Systems (NLDB 11), Springer, 2011, pp. 40-51.
[10] Abacha A, Zweigenbaum P. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies[J]. Information Processing and Management, 2015, 51:570-594.
[11] 孙辉, 雷枫. 中华人民共和国史本体构建初探[J].现代情报,2014,34(2):32-42.

作者贡献说明：

王颖：论文撰写、提出研究方案、系统详细设计和实现；
张智雄：提出研究思路，论文修订；
孙辉，雷枫：功能设计、数据加工。

Construction of Knowledge Retrieval Platform based on Historic Ontology of the People's Republic of China

Wang Ying[1] Zhang Zhixiong[1] Sun Hui[2] Lei Feng[2]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](The Institute of Contemporary China Studies, Beijing 100009, China)

Abstract: [Purpose/Significance] To build a historic knowledge retrieval platform, improve the efficiency access for users to history of the People's Republic of China, and promote its publicity and education. [Method/Process] This paper proposes the construction idea and framework of the knowledge retrieval platform based on historic ontology of the People's Republic of China. Based on the ontology knowledge base, this platform uses Neo4j database as data storage, creates three index based on Solr, including instance index, triple index and text item index. For various retrieval demands, the execution process of retrieval engine, construction method of retrieval expression, query processing algorithm and knowledge visualization are designed and implemented. [Result/Conclusion] The knowledge retrieval platform has been constructed, which provides entity search, query answering, relevance search, temporal retrieval and semantic resources browsing services. Its framework and implement of key technologies can provide an important reference for depth retrieval service on other domain knowledge.

Keywords: ontology entity search question answering relevance search visualization